UMIC: An Unreferenced Metric for Image Captioning via Contrastive Learning

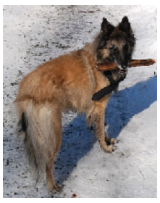
Hwanhee Lee[1], **Seunghyun Yoon**[2], **Franck Dernoncourt**[2]
Trung Bui[2] and **Kyomin Jung**[1]
[1]Dept. of Electrical and Computer Engineering, Seoul National University, Seoul, Korea
[2]Adobe Research, San Jose, CA, USA,
{wanted1007, kjung}@snu.ac.kr
{syoon, franck.dernoncourt, bui}@adobe.com

Abstract

Despite the success of various text generation metrics such as BERTScore, it is still difficult to evaluate the image captions without enough reference captions due to the diversity of the descriptions. In this paper, we introduce a new metric UMIC, an **U**nreferenced **M**etric for **I**mage **C**aptioning which does not require reference captions to evaluate image captions. Based on Vision-and-Language BERT, we train UMIC to discriminate negative captions via contrastive learning. Also, we observe critical problems of the previous benchmark dataset (i.e., human annotations) on image captioning metric, and introduce a new collection of human annotations on the generated captions. We validate UMIC on four datasets, including our new dataset, and show that UMIC has a higher correlation than all previous metrics that require multiple references. We release the benchmark dataset and pre-trained models to compute the UMIC[1].

1 Introduction

Image captioning is a task that aims to generate a description that explains the given image in a natural language. While there have been many advances for caption generation algorithms (Vinyals et al., 2015; Anderson et al., 2018) and target datasets (Fang et al., 2015; Sharma et al., 2018), few studies (Vedantam et al., 2015; Anderson et al., 2016; Cui et al., 2018; Lee et al., 2020) have focused on assessing the quality of the generated captions. Especially, most of the evaluation metrics only use reference captions to evaluate the caption although the main context is an image. However, as shown in Figure 1, since there are many possible reference captions for a single image, a candidate caption can receive completely different scores depending on the type of reference (Yi et al., 2020). Because of this diverse nature of image captions, reference-based metrics usually use multiple references which are difficult to obtain. To overcome this limitation, we propose UMIC, an **U**nreference **M**etric for **I**mage **C**aptioning, which is not dependent on the reference captions and use an image-caption pair to evaluate a caption. We develop UMIC upon UNITER (Chen et al., 2020) which is a state-of-the-arts pre-trained representation for vision-and-language tasks. Since UNITER is pre-trained to predict the alignment for large amounts of image-text pairs, we consider that UNITER can be a strong baseline for developing an unreferenced metric. We fine-tune UNITER via contrastive learning, where the model is trained to compare and discriminate the ground-truth captions and diverse synthetic negative samples. We carefully prepare the negative samples that can represent most of the undesirable cases in captioning, such as *grammatically incorrect*, *irrelevant to the image*, or *relevant but have wrong keyword*.



Ref 1: A dog standing in the snow with a stick in its mouth.
Ref 2: A little dog holding sticks in its mouth.
Candidate: A dog standing on the snow with a dog
CIDEr with Ref 1: 3.166
CIDEr with Ref 2: 0.281
Human Judgments : 1.875 out of 5

Figure 1: An example where the metric score for a given candidate caption varies significantly depending on the reference type.

When evaluating the metric's performance, it is required to compare the correlations between human judgments and the metric's evaluation score for given datasets. We choose three standard benchmark datasets (i.e., Composite (Aditya et al., 2015), Flickr8k (Hodosh et al., 2013), PASCAL-50s (Vedantam et al., 2015)) and further analyze the quality of the dataset. Interestingly, we found that there exist critical issues in the benchmark datasets,

[1]https://github.com/hwanheelee1993/UMIC

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Short Papers), pages 220–226
August 1–6, 2021.

such as poor-label or polarized-label. To perform a rigorous evaluation as well as stimulate the research in this area, we collect new 1,000 human judgments for the model-generated caption. Finally, we evaluate our proposed metric on four benchmark datasets, including our new dataset. Experimental results show that our proposed unreferenced metric is highly correlated with human judgments than all of the previous metrics that use reference captions.

2 Related Work

Image Captioning Metrics Following other text generation tasks such as dialogue systems and machine translation, n-gram similarity metrics such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004) and METEOR (Banerjee and Lavie, 2005) are widely used to evaluate an image caption. Especially, CIDEr (Vedantam et al., 2015), which weights each n-gram using TF-IDF, is widely used. SPICE (Anderson et al., 2016) is a captioning metric based on scene graph. BERTScore (Zhang et al., 2019), which computes the similarity of the contextualized embeddings, are also used. BERT-TBR (Yi et al., 2020) focuses on the variance in multiple hypothesis and ViLBERTScore (VBTScore) (Lee et al., 2020) utilizes ViLBERT (Lu et al., 2019) to improve BERTScore.

Different from these metrics, VIFIDEL (Madhyastha et al., 2019) computes the word mover distance (Kusner et al., 2015) between the object labels in the image and the candidate captions, and it does not require reference captions. Similar to VIFIDEL, our proposed UMIC does not utilize the reference captions. However, UMIC directly uses image features and evaluates a caption in various perspectives compared to VIFIDEL.

Quality Estimation Quality Estimation (QE) is a task that estimates the quality of the generated text without using the human references and this task is same as developing an unreferenced metric. QE is widely established in machine translation (MT) tasks (Specia et al., 2013; Martins et al., 2017; Specia et al., 2018). Recently, (Levinboim et al., 2021) introduces a large scale human ratings on image-caption pairs for training QE models in image captioning tasks. Our work also trains caption QE model, (i.e. unreferenced captioning metric) but we do not use human ratings to train the metric. Instead, we create diverse synthetic negative samples and train the metric with these samples via ranking loss.

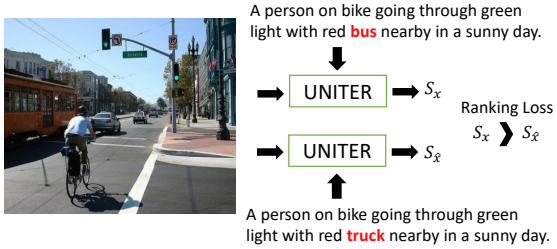


Figure 2: Overall training procedure of UMIC. Given an image I, a positive caption x and a negative caption $\hat{x}$, we compute the score of each image-caption pair S_x and $S_{\hat{x}}$ using UNITER respectively. Then, we fine-tune UNITER using raking loss that S_x is higher than $S_{\hat{x}}$.

3 UMIC

We propose UMIC, an unreferenced metric for image captioning using UNITER. We construct negative captions using the reference captions through the pre-defined rules. Then, we fine-tune UNITER to distinguish the reference captions and these synthetic negative captions to develop UMIC.

3.1 Modeling

Since UNITER is pre-trained to predict the alignment of large amounts of image-text pairs, we use the output of the layer that predicts this alignment as the baseline of UMIC to be fine-tuned. Specifically, we compute the score of a caption $\mathrm{S}(I, X)$ for given image $I = (i_1, ..., i_N)$ and $X = (x_1, ..., x_T)$ as follows.

We first compute the contextual embedding for I and X using UNITER to get the joint representation of image and text as follows.

$$i_{[CLS]}, i_1, ..., i_N, x_1, ..., x_T = \text{UNITER}(I, X), \quad (1)$$

where $i_{[CLS]}$ is a joint representation of the input image and input caption. Then we feed it into a single fully-connected layer to get a score as follows.

$$\mathrm{S}(I, X) = sigmoid(W i_{[CLS]} + b), \quad (2)$$

where W and b are trainable parameters.

3.2 Negative Samples

To model negative captions, we observe the captions' common error types in the model-generated captions. Specifically, we pick 100 bad captions in the order of whose human judgments are low in Composite and Flickr8k, respectively. Then, we categorize the main errors into three types:*relevant but have wrong keywords, totally irrelevant to the image, grammatically incorrect.* To model most

Target Image

Similar Image

Original: a woman hugging a girl who is holding a suitcase

Substitiution: a boy hugging a girl who is holding a suitcase

Random(Hard Negative): a very small cute child by a suitcase

Repetition & Removal: a woman hugging a girl is holding a suitcase suitcase

Figure 3: An example of the generated negative captions for the left image to train UMIC. Hard negative caption is one of the reference captions for the right image which is similar to the left image.

imperfect captions including these frequent type errors, we prepare negative captions as follows.

Substituting Keywords To mimic the captions that are relevant but have wrong keywords, as in the example of Figure 2, we randomly substitute 30% of the words in the reference captions and use them as negative samples like Figure 3. The motivation we choose 30% is that the average length of the generated caption is about 10 words and the number of keywords is usually around three. We only substitute *verb*, *adjective*, and *noun*, which are likely to be keywords since they are usually visual words. Also, we substitute them with the words with the same POS-Tags using the pre-defined dictionaries for the captions in the training set to conserve the sentence structure.

Random Captions We randomly sample captions from other images and use them as negative samples to generate totally irrelevant captions for the given image. Also, similar to the image-text retrieval task, we use hard-negative captions, which are difficult to be discerned, with a probability of 50%. Specifically, we utilize the captions of the images similar to the given images using the pre-trained image retrieval model. We get negative captions that are the captions of the similar image sets computed by image-text retrieval model VSE++ (Faghri et al., 2018) as in (Wang et al., 2020). Then, we sample the captions in the reference captions of the Top-3 similar image sets like the example in Figure 3.

Repetition & Removal We find that some of the captions have repeated words or have incomplete sentences. Hence, we randomly repeat or remove some words in the reference captions with a probability of 30% in the captions to generate these kinds of captions. Specifically, we choose to repeat or remove with a probability of 50% for the sampled word.

Word Order Permutation We further generate negative samples by randomly changing the word order of the reference captions, so that the model sees the overall structure of the sentence, not just the specific visual words.

3.3 Contrastive Learning

Using the negative captions generated by the above rules, we fine-tune UNITER via contrastive loss for positive caption X and negative caption $\hat{X}$ as follows.

$$Loss = max(0, M - (\mathrm{S}(I, X) - \mathrm{S}(I, \hat{X}))), \quad (3)$$

where M is the margin for the ranking loss, which is a hyperparameter. We make each batch composed of one positive caption and four negative captions that are made by each negative sample generation technique.

4 Dataset

We briefly explain the previous benchmark datasets for captioning metrics and analyze the problems for two of these datasets, Flickr8k and Composite. Also, we introduce a new benchmark dataset to alleviate the addressed problems.

4.1 Commonly Used Datasets

Composite consists of 11,985 human judgments for each candidate caption generated from three models and image pair. This dataset's human judgments range from 1 to 5, depending on the relevance between candidate caption and image.

Flickr8k provides three expert annotations for each image and candidate caption on 5,822 images. The score ranges from 1 to 4, depending on how well the caption and image match. All of the captions in this dataset are reference captions or captions from other images.

PASCAL50s contains 1,000 images from UIUC PASCAL Sentence Dataset with 50 reference captions for each image. Different from other datasets, this dataset provides 4,000 caption triplet $<A, B, C>$ composed of 50 reference captions(A) and two candidate captions(B, C) for the given image. There

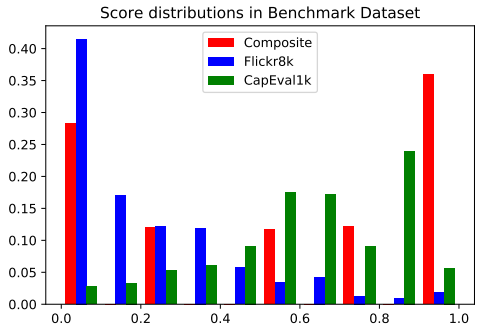


Figure 4: Score distributions of human judgments in Composite, Flickr8k and our proposed CapEval1k dataset. All scores were normalized from 0 to 1.

are human annotated answers to which is more similar to "*A*", "*B*" or "*C*".

4.2 Problems in Flickr8k and Composite

We investigate the human judgments in Flickr8k and Composite, and visualize the distributions of judgment scores for two datasets, Flickr8k and Composite in Figure 4, and find several problems.

For the Flickr8k, most of the scores are less than 0.2 since the candidate captions were sampled by an image retrieval system from a reference caption pool, not model-generated captions. Therefore, most captions are not related to images and differ significantly from the model-generated captions. We argue that this naive configuration is not enough to distinguish the performance of the metric precisely.

For the Composite, most of the scores are placed near 0 or 1. We explain this because only a single annotator annotates each sample's score resulting in biased output. We also manually investigated the captions and found that the captions are coarsely generated. Note that the captions for this dataset were generated by the old model (Karpathy and Fei-Fei, 2015; Aditya et al., 2015). For these reasons, we conclude that additional benchmark dataset is necessary to evaluate the captioning metrics.

4.3 CapEval1k Dataset

To alleviate the addressed issues in Flickr8k and Composite, we introduce a new dataset CapEval1k, which is composed of human judgments for the model-generated captions from four recently proposed models: Att2in (Rennie et al., 2017), Transformer (Vaswani et al., 2017), BUTD (Anderson et al., 2018) and AoANet (Huang et al., 2019). Different from Flickr8k and Composite, we ask each annotator to evaluate the captions by considering three dimensions: *fluency*, *relevance*, *descriptiveness*. We hire 5 workers who are fluent in English for each assignment from Amazon Mechanical Turk and use the average score. We also provide the full instructions and details in Appendix.

Metric	Flickr8k	Composite	CapEval1k	PASCAL50s
BLEU-1	0.274	0.406	0.233	74.3
BLEU-4	0.286	0.439	0.238	73.4
ROUGE-L	0.300	0.417	0.220	74.9
METEOR	0.403	0.466	0.288	78.5
CIDEr	0.419	0.473	0.307	76.1
SPICE	0.457	0.486	0.279	73.6
BERTScore	0.396	0.456	0.273	79.5
BERT-TBR	0.467	0.439	0.257	**80.1**
VBTScore	**0.525**	**0.514**	**0.352**	79.6
VIFIDEL	0.336	0.191	0.143	70.0
UMIC	**0.468**	**0.561**	**0.328**	**85.1**
UMIC$_C$	0.431	0.554	0.299	84.7

Table 1: Columns 1 to 3 represent Kendall Correlation between human judgments and various metrics on Flickr8k, Composite and CapEval1k. All p-values in the results are < 0.01. The last column shows the accuracy of matches between human judgments in PASCAL50s.

Since our CapEval1k dataset is composed of annotations via recently proposed models, the overall scores are relatively higher than other datasets as shown in Figure 4. Compared to other datasets, CapEval1k contains the annotators' comprehensive judgment across multiple dimensions in evaluating the quality of the generated captions, so we can see that the score distribution score is not concentrated in a particular area.

5 Experiments

5.1 Implementation Details

We use the pre-trained UNITER-base with 12 layers in the official code provided by the authors (Chen et al., 2020)[2]. We use the COCO dataset (Fang et al., 2015) to fine-tune UNITER through ranking loss. We use the train and validation split of COCO dataset in (Chen et al., 2020). The number of the training set is 414k, and the validation set is 25k. We set the batch size of 320, learning rate of 2e-6, and fine-tune UNITER for a maximum of 4k steps. We select the model that shows the minimum loss in the validation set. We set margin M as 0.2 in the ranking loss. We repeat training 5 times for each best-performing model.

5.2 Performance Comparison

We compute caption-level Kendall's correlation coefficient with human judgments for the Composite,

[2]https://github.com/ChenRocks/UNITER

Flickr8k, and our proposed CapEval1k. For the PASCAL50s, we compute the number of matches between human judgments for each candidate caption pair. For all of the reference based metrics, we use five reference captions and then get average score among the five references except for BERTScore where we use maximum.

We present the experimental results for all four datasets in Table 1. We show that although UMIC does not utilize any reference captions, UMIC outperforms the baseline metrics except for VBTScore in all of the datasets that depend on multiple references. We also report the strong unreferenced baseline UMIC$_{\text{-C}}$, which is directly using the pre-trained weights from UNITER without contrastive learning. Interestingly, UMIC$_{\text{-C}}$ shows a higher performance than most of the metrics. This high performance shows that pre-trained image-text matching layer of UNITER already has a good representation for evaluating image captions. Especially for Composite, both UMIC and UMIC$_{\text{-C}}$ significantly outperform baseline metrics. We explain this in the polarized distribution of human judgments as we explained in Section 4.2. In other words, the relevance of most image-caption pairs in this dataset is too obvious so that UNITER can easily distinguish them. However, while UMIC shows higher performance on all datasets, UMIC$_{\text{-C}}$ shows relatively low performance on Flickr8k and CapEval1k. And this demonstrates the effectiveness and generalization ability of our contrastive learning objective to develop UMIC.

Also, we can observe that the performance of each metric is relatively low and the rank of each metric changes in our proposed CapEval1k dataset. We explain that this is because the captions in CapEval1k are relatively difficult to be evaluated since the score distribution is not biased as explained in Section 4.3.

5.3 Case Study

We visualize one sample each showing the strengths and weaknesses of UMIC in Figure 5. In the above example, the candidate caption is partially relevant to the image, but the single word "three" in the caption is totally incorrect since there are only "two" giraffes in the image. And this leads to a low human judgment of 0.2. Nevertheless, unlike our UMIC, widely used metrics and UMIC$_{\text{-C}}$ give this caption a high score due to the many words overlaps or missing the keywords. The bottom example shows one of the error cases and the limitations of our proposed method. Since the detection model in UMIC could not recognize the important object like the "baseball bat", UMIC outputs very low score.

References
- two giraffe standing next to each other in a field.
- two giraffes are climbing a hill with mountains in the background.

Candidate
- **three** giraffes standing in a field of grass

BLEU1: 0.324	**ROUGE-L**: 0.320	**METEOR**: 0.173	**CIDER**: 0.866
SPICE: 0.289	**UMIC**: 0.352	**UMIC**$_{/-C}$: 0.770	**Human**: 0.200

References
- a person breadking a bottle with a baseball bat
- a boy in yellow shirt swinging a baseball bat

Candidate
- a man swinging a **baseball bat** at a ball

BLEU1: 0.360	**ROUGE-L**: 0.354	**METEOR**: 0.176	**CIDER**: 1.205
SPICE: 0.192	**UMIC**: 0.094	**UMIC**$_{/-C}$: 0.062	**Human**: 0.450

Figure 5: Case study for the various metrics on candidate captions in CapEval1k Dataset. Human judgments are normarlized from 0 to 1.

6 Conclusion

In this paper, we propose UMIC, an unreferened metric that does not require any reference captions for image captioning task through contrastive learning in UNITER. Also, we propose a new benchmark dataset for image captioning that relieve the issues in previous datasets. Experimental results on four benchmark datasets, including our new dataset, show that UMIC outperforms previous metrics.

Acknowledgements

We thank anonymous reviewers for their constructive and insightful comments. K. Jung is with ASRI, Seoul National University, Korea. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (No. 2021R1A2C2008855). This work was partially funded by gifts from Adobe Research.

Ethical Considerations

We compensate the annotators with competitive pay, which is above hourly USD $10 for collecting human annotated judgments for the model generated captions. Specifically, we pay $0.2 for each task that is composed of evaluating four candidate captions for a single image, where each task can be usually done in a minute. And we use public datasets to train the models.

References

Somak Aditya, Yezhou Yang, Chitta Baral, Cornelia Fermuller, and Yiannis Aloimonos. 2015. From images to sentences through scene description graphs using commonsense reasoning and knowledge. *arXiv preprint arXiv:1511.03292*.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *European Conference on Computer Vision*, pages 382–398. Springer.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *European Conference on Computer Vision*, pages 104–120. Springer.

Yin Cui, Guandao Yang, Andreas Veit, Xun Huang, and Serge Belongie. 2018. Learning to evaluate image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5804–5812.

Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2018. Vse++: Improving visual-semantic embeddings with hard negatives.

Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. 2015. From captions to visual concepts and back. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1473–1482.

Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899.

Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. 2019. Attention on attention for image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4634–4643.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *International conference on machine learning*, pages 957–966.

Hwanhee Lee, Seunghyun Yoon, Franck Dernoncourt, Doo Soon Kim, Trung Bui, and Kyomin Jung. 2020. Vilbertscore: Evaluating image caption using vision-and-language bert. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 34–39.

Tomer Levinboim, Ashish V. Thapliyal, Piyush Sharma, and Radu Soricut. 2021. Quality estimation for image captions based on large-scale human evaluations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3157–3166, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pages 13–23.

Pranava Swaroop Madhyastha, Josiah Wang, and Lucia Specia. 2019. Vifidel: Evaluating the visual fidelity of image descriptions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6539–6550.

André FT Martins, Marcin Junczys-Dowmunt, Fabio N Kepler, Ramón Astudillo, Chris Hokamp, and Roman Grundkiewicz. 2017. Pushing the limits of translation quality estimation. *Transactions of the Association for Computational Linguistics*, 5:205–218.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7008–7024.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565.

Lucia Specia, Frédéric Blain, Varvara Logacheva, Ramón Astudillo, and André FT Martins. 2018. Findings of the wmt 2018 shared task on quality estimation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 689–709.

Lucia Specia, Kashif Shah, José GC De Souza, and Trevor Cohn. 2013. Quest-a translation quality estimation framework. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 79–84.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164.

Jiuniu Wang, Wenjia Xu, Qingzhong Wang, and Antoni B Chan. 2020. Compare and reweight: Distinctive image captioning using similar images sets. In *ECCV*.

Yanzhi Yi, Hangyu Deng, and Jinglu Hu. 2020. Improving image captioning evaluation by considering inter references variance. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 985–994.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.